Searchable Patent Embeddings

Aaron Guo, Lawrence Liu, Joyce Qu

Motivation

Status Quo

- A typical US Patent and Trademark Office (USPTO) search returns 2,000 results which are scarcely relevant. The top 20 patents from the prompt “garden AND hose” include a rotating lawn care contraption and an evaporator device for distillation.

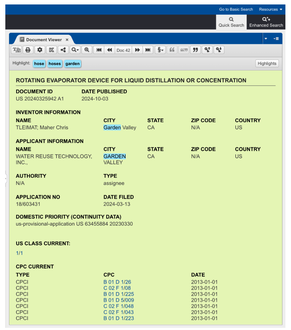


Fig. 1 USPTO search results “garden AND hose”

Patent Landscaping

- Process of using similarity to create a set of patents that span a particular topic

Goal

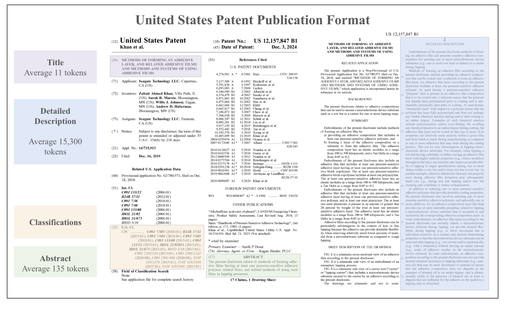
- Improve patent landscaping with embeddings such that we can build a patent landscape from a *single* patent

Significance of Improvement

- Prospective inventors avoid skimming under-focused results which take time and are not comprehensive.

Data

Corpus Format



Labels

Code	CPC Category
A	Human Necessities
B	Performing Operations; Transporting
C	Chemistry, Metallurgy
D	Textiles, Paper
E	Fixed Constructions
F	Mechanical Engineering; Lighting; Heating; Weapons; Blasting
G	Physics
H	Electricity
Y	General Tagging of New Technological Developments

Patent Universe: Random sample of 30,000 patents published in 2023

Architecture

Embeddings *TitleAbstract*, *Fulltext* Truncated, *Fulltext*, and *TitleAbstract* (Contextual Document Embeddings)

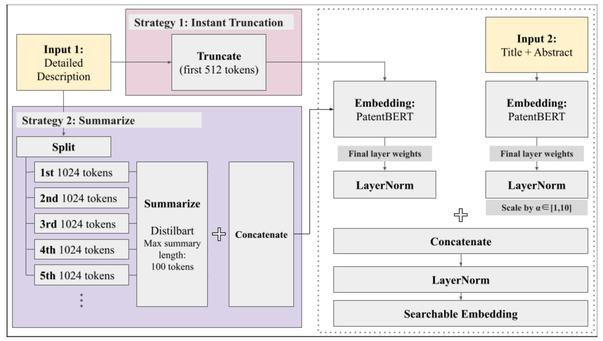


Fig. 2 *Fulltext* (Summarized and Truncated)

Evaluation Metrics

Classification Metrics

One versus many linear SVM that applies binary labels of CPC category to each of the document embeddings.

Similarity metrics

Closeness: A measure of the average proportion of same category patents that are in the top half of similarity.

$$A(p_i) = \frac{1}{|P(c(p_i))|} \sum_{p' \in P(c(p_i))} \mathbb{1}\left(s'_i[p'] < \frac{n}{2}\right)$$

$$Closeness = \frac{1}{n}\sum_{i=1}^{n} A(p_i)$$

Same category similarity (SCS): A measure of the proportion of patents that have a majority of same category category patents in the top half of similarity rankings.

$$SCS = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(A(p_i) > \frac{1}{2}\right)$$

Conclusions

Fulltext offers valuable information

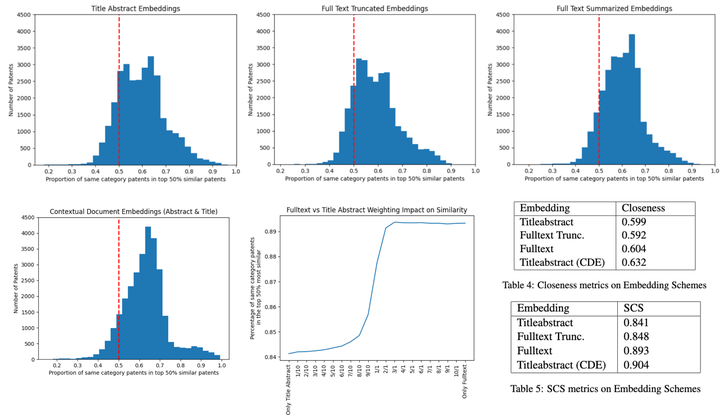
Fulltext embeddings outperform *truncated fulltext* embeddings, showing the viability of text summary in patent landscaping. It also mitigates gaps in existing literature that truncate patent full text.

Fulltext embeddings outperform exclusively title and abstract-based embeddings, and these descriptions offer a rich context for the embedding space.

Embeddings using patentBERT are close to or past the mark in tasks of similarity search and classification compared to state-of-the-art document encoders.

Results

Landscaping



Embedding	Closeness
Titleabstract	0.599
Fulltext Trunc.	0.592
Fulltext	0.604
Titleabstract (CDE)	0.632

Table 4: Closeness metrics on Embedding Schemes

Embedding	SCS
Titleabstract	0.841
Fulltext Trunc.	0.848
Fulltext	0.893
Titleabstract (CDE)	0.904

Table 5: SCS metrics on Embedding Schemes

Fig. 3 *Closeness* Distributions & *SCS*

Classification

Embedding	F1-score Sample Avg
Fulltext	0.79
Fulltext Truncated	0.78
Titleabstract	0.78
Titleabstract (CDE)	0.67

Table 3: Average F1-Scores for Embedding Schemes

section	precision	recall	f1-score	support
A	0.83	0.77	0.80	701
B	0.71	0.64	0.67	848
C	0.80	0.82	0.81	371
D	0.74	0.50	0.60	28
E	0.77	0.65	0.71	229
F	0.82	0.74	0.78	605
G	0.82	0.86	0.84	2564
H	0.85	0.81	0.83	2249
Y	0.00	0.00	0.00	20
None	0.85	0.03	0.05	667
Samples Avg	0.83	0.80	0.79	8282

Figure 4: Fulltext Classification

Fig. 4 Classifier Capacity & *Fulltext* in-depth breakdown

Interface Prototype

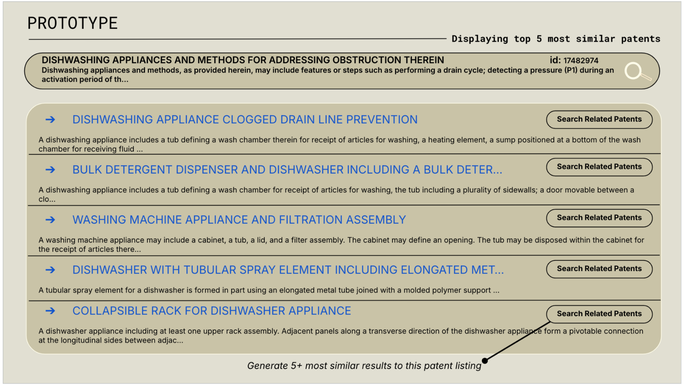


Fig. 5 Patent search interface with the *FullText* model showing top 5 results